

# Estimating the dimensionality of neural responses with fMRI Repetition Suppression

Mattia Rigotti<sup>1</sup> and Stefano Fusi<sup>2</sup>

<sup>1</sup> IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA

<sup>2</sup> Center for Theoretical Neuroscience, Department of Neuroscience, Columbia University Medical Center, New York, NY 10032, USA

**Abstract.** We propose a novel method that exploits fMRI Repetition Suppression (RS-fMRI) to measure the dimensionality of the set response vectors, i.e. the dimension of the space of linear combinations of neural population activity patterns in response to specific task conditions. RS-fMRI measures the overlap between response vectors even in brain areas displaying no discernible average differential BOLD signal. We show how this property can be used to estimate the neural response dimensionality in areas lacking macroscopic spatial patterning. The importance of dimensionality derives from how it relates to a neural circuit's functionality. As we show, the dimensionality of the response vectors is predicted to be high in areas involved in multi-stream integration, while it is low in areas where inputs from independent sources do not interact or merely overlap linearly. Our method can be used to identify and functionally characterize cortical circuits that integrate multiple independent information pathways.

## 1 Introduction

The pioneering work by Hubel and Wiesel [14] on the receptive field properties of neurons in early visual cortex has helped shape the remarkably influential notion that sensory information in the brain is processed in hierarchically organized computational modules. Feedforward stages where features are distinctly extracted and analyzed feed into alternating higher order convergent stages where such features are combined to form the increasingly complex receptive fields of more specialized cells [7].

The idea to use multiple hierarchical elaboration stages of increasing levels of abstraction has been embraced by the deep learning community with spectacular results in processing tasks of images, audio and video [17]. However, the general question as to how higher order features should be derived from the convergence of more basic ones is still unresolved, and has long been at the center of a multidisciplinary interest that spans computational and systems neuroscience, computer vision, and machine learning in general. One central idea from this interdisciplinary cross-pollination is that higher order features have to strike a balance between *dimensionality expansion* and *dimensionality reduction*. The convergence of lower order features within higher order modules has to

be non-linear, so as to guarantee dimensionality expansion and therefore high-margin separation [6,5]. On the other hand, non-linear dimensionality expansion is usually alternated with a dimensionality reduction step, such as max-pooling, which allows for better generalization by enforcing invariance to input changes that shouldn't affect the final processing output [10,2,25]. From a kernel machine perspective, good resulting high order features should be rich enough to serve as a "basis" to express the final output, while at the same time allowing for a parsimonious expression in terms of a reduced number of components [19]. This complements the emerging view in the deep learning community that in the higher layers of deep neural networks trained to perform classification tasks, the important semantic information is contained in the space of linear combinations of responses, rather than in the activity of individual units [24].

Going beyond sensory processing tasks, recent neuroscience studies have confirmed the importance of the specific structure of the linear space generated by the neural responses also during the execution of cognitive operations. In particular, the dimensionality of the response vectors in convergent frontal cortical areas has been shown to predict the behavior of rodents [4] and monkeys [21] during working memory tasks, and to collapse during error trials.

In this contribution we propose a way to quantify the dimensionality of the neural response patterns using repetition suppression fMRI (RS-fMRI). RS-fMRI or fMRI-adaptation [12,13] refers to a progressive reduction of BOLD response following repeated stimulus presentations. The subsequent presentation of differing stimuli results in a BOLD signal suppression that can be used to estimate the extent to which their representations overlap at the neural level. We propose to use this approach to determine how dissimilar responses are within a single voxel. This approach enables us to measure the relative distances between neural responses and produce an estimate of the dimensionality within a voxel.

As we will show in the Results section, this method can be used to investigate the role of specific brain regions in dimensionality expansion or reduction, potentially contributing to the understanding of the computational and algorithmic basis of higher order cognitive functions. As an application, we will illustrate the result of simulations indicating that our method can be used to identify and functionally characterize areas that integrate multiple information pathways.

## 2 Methods

### 2.1 Model of repetition suppression in fMRI

This section illustrates in detail the assumptions of our RS-fMRI model, and demonstrates how this technique can be leveraged to estimate the dimensionality of the response vectors within voxels.

**Additive model of fMRI BOLD signal from neural activity.** We denote with  $\mathbf{f}(e)$  the response vector elicited in response to a task-relevant event  $e$  within a brain region or voxel of interest of  $N$  neurons. Accordingly, the firing rate activity of neuron  $i$  during event  $e$  is indicated by  $f_i(e)$ . We are interested in

the dimensionality of the space spanned by the vectors  $\mathbf{f}(e_k)$ ,  $k = 1, \dots, n$ . This quantity equals the rank of the matrix  $F = (\mathbf{f}(e_1), \mathbf{f}(e_2), \dots, \mathbf{f}(e_n))$  obtained by concatenating all response vectors.

We model the fMRI signal in the region of interest as an aggregate function of the response function of all  $N$  neurons. In particular, we assume that the BOLD signal evoked by event  $e$  follows the simple form:

$$B_e = \beta \sum_i^N f_i(e), \quad (1)$$

where  $\beta$  is a positive constant relating neural activation to BOLD signal. Notice that  $B_e$  is a function of the population response vector evoked by event  $e$ , but is a scalar. Previous studies coupling electrode recordings and intrinsic optical imaging of BOLD signals demonstrate the plausibility of assuming such a linear relation between BOLD and spiking activity [8]. In all simulations we add a Gaussian random variable to  $B_e$  to simulate fMRI noise [15].

**Firing rate adaptation model of RS-fMRI.** We mechanistically model repetition suppression as being due to firing rate adaptation or “neuronal fatigue” [12,9]. We set out to model the neural response elicited by the presentation of sequences of two events  $e_1$  and  $e_2$  in a fixed temporal order. In particular, we model the fact that the response vector at the presentation of the second event will be modified, due to the previous activation at the first event. Our model assumes that the firing rate response of neuron  $i$  at the first event  $e_1$  is given by  $f_i(e_1) = \phi(I_i(e_1))$ , where  $I_i(e_1)$  indicates the total synaptic current to neuron  $i$  and  $\phi(\cdot)$  indicates the neuronal current-to-firing rate transfer function.

We will use the model of firing rate adaptation developed by [16], and assume that the firing response activates an hyperpolarizing current proportional to the response itself. More specifically, at the presentation of a second event  $e_2$ , following a fixed delay after event  $e_1$ , we model the response of neuron  $i$  as

$$f_i(e_1 e_2) = \phi(I_i(e_2) - \alpha_t f_i(e_1)), \quad (2)$$

where  $I_i(e_2)$  is the synaptic drive elicited by event  $e_2$  onto neuron  $i$ ,  $\alpha_t$  is a positive parameter that quantifies the adaptation strength (the subscript  $t$  denotes the dependence of the adaptation effect to the time interval between  $e_1$  and  $e_2$ ), and  $f_i(e_1)$  is the unadapted response to event  $e_1$ .

**RS-fMRI and neural response vectors dimensionality.** As it is often done, we assume that the neurons have a transfer function that goes asymptotically to zero for decreasing inputs and tends asymptotically to a linear function for growing inputs [1]. In-between, we assume that the transfer function can be modeled as an exponential:  $\phi(I) = \exp(I/\sigma + b)$ , with some constant  $\sigma$  and a bias term  $b$ . In this regime, equation (2) assumes the multiplicative form:

$$f_i(e_1 e_2) = \phi(I_i(e_2) - \alpha_t f_i(e_1)) = \rho(f_i(e_1)) f_i(e_2),$$

with  $\rho(x) = \exp(-\alpha_t x/\sigma)$ , resulting in a fMRI-RS signal described by:

$$B_{e_1 e_2} = \beta \sum_i \rho(f_i(e_1)) f_i(e_2). \quad (3)$$

## IV

Notice that the factors  $\rho(f_i(e_1))$  are implicitly also a function of the inter-event interval between  $e_1$  and  $e_2$  through the dependence on  $\alpha_t$ . In the limit where the two events are very distant in time, their value should in fact tend to 1, and  $B_{e_1 e_2}$  tend to  $B_{e_2}$ . In that limit, we can expand  $\rho(\cdot)$  to first order as  $\rho(f) \approx 1 - \alpha_t f / \sigma$ . We now define the Repetitions Suppression (RS) matrix with components  $R_{kl}$ :

$$R_{kl} = B_{e_l} - B_{e_k e_l} \quad (4)$$

and notice that plugging the approximation  $\rho(f) = 1 - \alpha_t f / \sigma$  into eq. (3) together with eq. (1) allows us to derive

$$R_{kl} = B_{e_l} - B_{e_k e_l} \propto \sum_i^N f_i(e_k) f_i(e_l) = (F^\top F)_{kl},$$

where  $F$  is the previously defined matrix of response vectors. Now, because in general  $\text{rank}(F) = \text{rank}(F^\top F)$ , we get that the rank of the response vectors  $F$  is the same as the rank of the RS-matrix  $R_{kl}$ .

Going through the same calculation in the case of a linear transfer function, it is easy to see that the constant and the asymptotically linear regime of the transfer function does not contribute to the rank. So, all contributions are given by the non-linear exponential part. Notice also that a threshold linear function (the transfer function we will use in the simulation) can be approximated by a piece-wise exponential and linear function.

As a final note it is important to mention that the estimate of the rank of a matrix is made notoriously difficult by random noise fluctuations. These random variations bias the estimate of the principal components, increasing the number of non-zero components that contribute to the rank (see e.g. [23]).

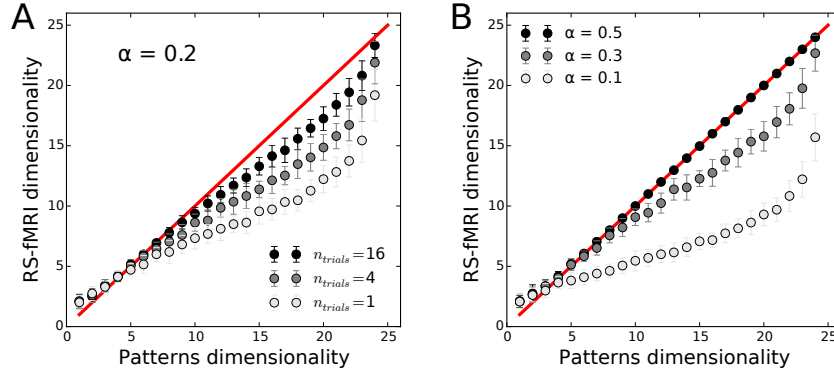
We determine the noise floor similarly to [18] by estimating a noise covariance matrix from single-trial data that we use to build a distribution of “noise eigenvalues”. We then set the noise threshold to 95% of the maximal noise eigenvalue. In all simulations we set a noise level that corresponds to a SNR of 100, a value that matches measures of physiological noise in gray matter [15].

## 3 Results

### 3.1 RS-matrix estimate of dimensionality of response vectors

The following simulations validate our method by showing that our model of BOLD response eq. (1) with adapting neurons eq. (2) allows us to reliably estimate the dimensionality of the response vectors from the RS-matrix  $R$  in eq. (4). We generate sets of  $p$   $N$ -dimensional patterns with varying rank by multiplying an  $N \times d$ -dimensional random Gaussian matrix (centered at zero and with unitary variance) with a  $d \times p$  random binary matrix. These vectors correspond to the synaptic currents  $I_i(e)$  generating the firing rates  $f_i(e) = \phi(I_i(e))$ . As a transfer function  $\phi(\cdot)$  we use a threshold-linear function  $\phi(I) = \max(I, 0)$ . We then use eq. (2) with eq. (4) and (1) to generate the RS-matrix  $R$ . For numerical stability

reasons, we then symmetrize this matrix by averaging it with its own transpose. We then estimate the dimensionality by comparing the eigenvalues of this matrix to the noise threshold as explained in the Methods section and compare it with the dimensionality obtained from the matrix  $F_{ik} = f_i(e_k)$ . Figure 1 shows the result of such a simulation, demonstrating that the estimated dimensionality accurately correlates with the actual response vectors dimensionality.

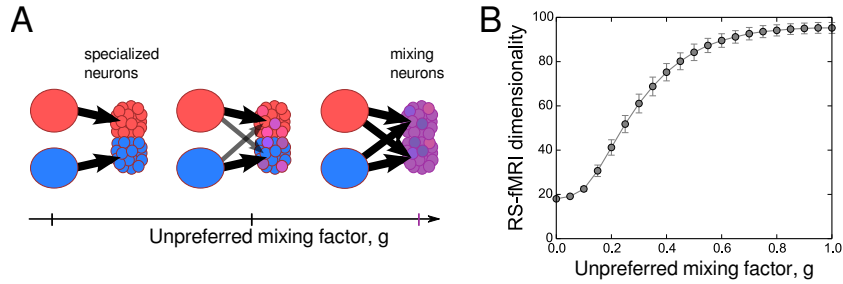


**Fig. 1.** Estimate obtained from the RS-matrix accurately quantifies the dimensionality of the neural response vectors. **A** Dimensionality estimated from the RS-matrix is highly correlated with the actual patterns dimensionality. Moreover, the correlation grows as more and more trials are used to average out noise in the RS-fMRI estimate. The plot shows the RS-fMRI estimated dimensionality as a function of actual dimensionality with no trial-to-trial averaging (white circles), average over 4 trials (gray circles) and average over 16 trials (black circles) for a fixed number of  $p = 25$  patterns with rank growing between 1 and 25. The circles indicate the mean over 100 simulations, while error bars denote the corresponding standard deviations. **B** As in **A**, but with no trial-to-trial average (only 1 trial) and simulating different values of the adaptation factor  $\alpha$ .

### 3.2 Functional dissection of information processing pathways

We now illustrate how our technique could be used for the functional dissection of neural information processing in the brain. We imagine a neural circuit that receives synaptic projections from two separate independent input areas  $A$  and  $B$ , each of which can be in 10 states indexed by  $k_A$  and  $k_B$ , respectively. In our model every neuron has a preferred input (either  $A$  or  $B$ ) from which it receives most of its inputs. So, a neuron  $i$  with a preference for  $A$  will receive a current  $I_i(k_A, k_B) = \sum_j w_{ij}^A f_j^A(k_A) + g \sum_j w_{ij}^B f_j^B(k_B)$ , where the “unpreferred mixing factor”  $g$  is a number from 0 and 1 that quantifies the contribution of the unpreferred input,  $w^A$  and  $w^B$  are random centered Gaussian matrices with unitary variance, while  $f_j^A(k_A)$  and  $f_j^B(k_B)$  are the activities of the neurons

in the two input areas (also taken as centered random Gaussian variables with unitary variance in our simulations). Everything works analogously for neurons with a  $B$  preference. With  $g = 0$  the neurons in the convergence area will only receive projections from their preferred input, while for  $g = 1$  the neurons will equally mix contribution from both inputs (see Fig. 2A). Panel Fig. 2B shows the results of the simulation carried out similarly to the one in the previous section to simulate the process of estimating the dimensionality of the responses in the convergence areas. If the neurons in the convergence area only receive input from their preferred area the dimensionality is estimated to be  $19 = 20 - 1$ , the result of separately counting the contribution of the two areas (see [5]). As the mixing factor increases and neurons mix more and more the contributions from the two inputs, the measured dimensionality increases until it saturates around  $99 = 100 - 1 = 10 \times 10 - 1$  when the two inputs are equally mixed (see [5]). This could be used to identify areas that receive contributions from multiple independent inputs, and to characterize the type of functional convergence.



**Fig. 2.** Functional dissection of convergent pathways with RS-fMRI. **A** Scheme of the simulated circuit. The input areas project to a convergence area. The way the inputs are mixed is parametrized by  $g$  (between 0 and 1). For  $g = 0$  every neuron in the convergence area receives inputs only from its preferred input area. As  $g$  increases, the input from the unpreferred area increases until  $g = 1$  where the neurons mix on average equal contributions from both input areas. **B** Measuring the dimensionality of the response vectors in the convergence area as in Fig. 1 shows, as expected (see [5]), that the dimensionality grows as the mixing of the two inputs increases.

## 4 Discussion

We proposed a novel technique for designing experiments and analyzing data that will allow us to estimate dimensionality in single voxels using fMRI. If our mechanistic model is accurate, the dimensionality estimate given by our method should match the result from single unit recordings, in the sense that having access to the individual activity of all neurons in the voxel wouldn't give any advantage. This is because of the mathematical identity relating the

rank of the response vectors and that of their overlaps captured by the RS-matrix. Here we put forth the main idea with analytical calculations on realistic mechanistic models of repetition suppression and we showed two simulations illustrating possible applications. Further simulations (omitted because of space constraints) show that our technique is robust to variations of the RS model (e.g. the inclusion of synaptic effects such as “input fatigue”), to different choices of non-linear transfer functions, forms of noise, and heterogeneity of RS parameters.

Further ongoing studies will compare the efficacy of this new technique to alternative methods for estimating dimensionality based on Multi-Voxel Pattern Analysis (MVPA) [20]. Crucially, MVPA relies on a differential signal within individual voxels, while voxels whose BOLD activity is not modulated across task conditions, won’t contribute to decoding. It is important to note that this situation where a voxel does not contribute to the differential BOLD signal can arise in two distinct mechanistic cases: 1) the trivial scenario where the activity of individual neurons within the voxel don’t display any selectivity to task conditions, and 2) the case where neurons are indeed selective to specific conditions, but the preferences across neurons within the voxel are balanced in such a way that the signal is averaged out when it is aggregated in the total BOLD activity. In the second case, RS-fMRI can still distinguish between conditions, since the activity patterns are effectively distinct.

We therefore expect MVPA to be maximally effective when voxels are organized in such a way as to display a net preference, due to an underlying preference similarity (which can be quantified as signal correlation [3]) across its constituting neurons, while we expect it to be progressively less effective as the preference of neurons in the voxels is less and less correlated, until the extreme case where it completely fails due to preferences being perfectly balanced. Notice that the situation in which we are interested is one where no particular anatomical organization is present, such as it is observed in higher order cortices, and therefore where we expect RS-fMRI to have a competitive advantage over MVPA.

These intuitive considerations are borne out quantitatively by a model of anatomical organization that shows that in the limit of disappearing signal correlation between neurons in a voxel, the MVPA signal tends to a small term that scales as  $1/\sqrt{N}$ , where  $N$  is the number of neurons in the voxel (a finite-size effect signal), while the RS-fMRI signal remains of the same order as with finite signal correlation [22].

Clearly, our method remains to be demonstrated empirically. In this respect we’re encouraged by previous studies showing that RS-fMRI can be utilized to distinguish between cortical areas tuned to either independent or conjoint visual features [11]. A conclusive empirical verification of our technique for the measure of response dimensionality is however left to future experiments.

## 5 Acknowledgements

SF was supported by the Gatsby Charitable Foundation, the Simons Foundation, the Swartz Foundation, the Kavli Foundation and the Grossman Foundation.

## References

1. Abbott, L., Chance, F.S.: Drivers and modulators from push-pull and balanced synaptic input. *Progress in brain research* 149, 147–155 (2005)
2. Anselmi, F., Leibo, J.Z., Rosasco, L., Mutch, J., Tacchetti, A., Poggio, T.: Unsupervised learning of invariant representations in hierarchical architectures. *arXiv preprint arXiv:1311.4158* (2013)
3. Averbeck, B.B., Latham, P.E., Pouget, A.: Neural correlations, population coding and computation. *Nat Rev Neurosci* 7(5), 358–366 (May 2006), <http://dx.doi.org/10.1038/nrn1888>
4. Balaguer-Ballester, E., Lapish, C.C., Seamans, J.K., Durstewitz, D.: Attracting dynamics of frontal cortex ensembles during memory-guided decision-making. *PLoS Comput Biol* 7(5), e1002057 (May 2011), <http://dx.doi.org/10.1371/journal.pcbi.1002057>
5. Barak, O., Rigotti, M., Fusi, S.: The sparseness of mixed selectivity neurons controls the generalization-discrimination trade off. *Journal of Neuroscience* 33(6), 3844–3856 (Feb 2013)
6. Barak, O., Rigotti, M.: A simple derivation of a bound on the perceptron margin using singular value decomposition. *Neural Comput* 23(8), 1935–1943 (Aug 2011), [http://dx.doi.org/10.1162/NECO\\_a\\_00152](http://dx.doi.org/10.1162/NECO_a_00152)
7. Barlow, H.B.: Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* 1(4), 371–394 (1972)
8. Cardoso, M.M.B., Sirotin, Y.B., Lima, B., Glushenkova, E., Das, A.: The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. *Nature Neuroscience* 15(9), 1298–1306 (Sep 2012), <http://dx.doi.org/10.1038/nn.3170>
9. De Baene, W., Vogels, R.: Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. *Cereb Cortex* 20(9), 2145–2165 (Sep 2010), <http://dx.doi.org/10.1093/cercor/bhp277>
10. DiCarlo, J.J., Cox, D.D.: Untangling invariant object recognition. *Trends Cogn Sci* 11(8), 333–341 (Aug 2007), <http://dx.doi.org/10.1016/j.tics.2007.06.010>
11. Drucker, D.M., Kerr, W.T., Aguirre, G.K.: Distinguishing conjoint and independent neural tuning for stimulus features with fmri adaptation. *J Neurophysiol* 101(6), 3310–3324 (Jun 2009), <http://dx.doi.org/10.1152/jn.91306.2008>
12. Grill-Spector, K., Henson, R., Martin, A.: Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences* 10(1), 14–23 (2006)
13. Grill-Spector, K., Malach, R.: fmr-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta psychologica* 107(1), 293–321 (2001)
14. Hubel, D. H., D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat’s striate cortex. *J Physiol* 148, 574–591 (Oct 1959)
15. Krüger, G., Glover, G.H.: Physiological noise in oxygenation-sensitive magnetic resonance imaging. *Magnetic Resonance in Medicine* 46(4), 631–637 (2001)
16. La Camera, G., Rauch, A., Lüscher, H.R., Senn, W., Fusi, S.: Minimal models of adapted neuronal response to in vivo-like input currents. *Neural Comput* 16(10), 2101–2124 (Oct 2004), <http://dx.doi.org/10.1162/0899766041732468>
17. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* 521(7553), 436–444 (May 2015), <http://dx.doi.org/10.1038/nature14539>
18. Machens, C.K., Romo, R., Brody, C.D.: Functional, but not anatomical, separation of “what” and “when” in prefrontal cortex. *J Neurosci* 30(1), 350–360 (Jan 2010), <http://dx.doi.org/10.1523/JNEUROSCI.3276-09.2010>



19. Montavon, G., Braun, M.L., Müller, K.R.: Kernel analysis of deep networks. *The Journal of Machine Learning Research* 12, 2563–2581 (2011)
20. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fmri data. *Trends Cogn Sci* 10(9), 424–430 (Sep 2006), <http://dx.doi.org/10.1016/j.tics.2006.07.005>
21. Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., Fusi, S.: The importance of mixed selectivity in complex cognitive tasks. *Nature* 497(7451), 585–590 (May 2013), <http://dx.doi.org/10.1038/nature12160>
22. Rigotti, M., Fusi, S.: In preparation
23. Sengupta, A.M., Mitra, P.P.: Distributions of singular values for some random matrices. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 60(3), 3389–3392 (Sep 1999)
24. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013)
25. Wang, W., Carreira-Perpinán, M.A.: The role of dimensionality reduction in linear classification. *arXiv preprint arXiv:1405.6444* (2014)